Validation of AI-Enabled Surrogate Models in in Quantitative Systems Pharmacology

A Practical, Context-of-Use–Driven Framework

Igor Goryanin[*1,2], *Stephen Checkley*[2], *& Irina Goryanin*[2]

1. Artificial Intelligence Institute, School of Informatics, University of Edinburgh · 2. IQANOVA Ltd., Edinburgh

Abstract

The Challenge

QSP models are computationally intensive—virtual population generation, global sensitivity analysis, and Bayesian parameter estimation require thousands to millions of evaluations. AI-enabled surrogates offer orders-of-magnitude speed-ups but lack structured validation aligned with regulatory expectations.

This Framework

An eight-step, end-to-end validation workflow aligned with risk-informed credibility standards. Covers surrogate classes (Gaussian processes, deep neural networks, hybrid mechanistic–ML), validation strategies for endpoint accuracy, trajectory fidelity, uncertainty calibration, and virtual population distributional agreement, plus failure-mode diagnostics.

Case Studies & Scope

CAR-T immunotherapy and a hybrid Neural ODE erythropoiesis surrogate illustrate domain-specific requirements. Regulatory implications are discussed under FDA CM&S and ASME V&V40 frameworks.

Why Surrogate Validation Is Different in QSP

QSP integrates mechanistic biological modelling with PK/PD to support translational research and model-informed drug development across immuno-oncology, rare diseases, and metabolic disorders. Their underlying differential equations are nonlinear, high-dimensional, and numerically stiff—making tasks like GSA, VP generation, and Bayesian estimation computationally infeasible without surrogates.

Not Just Accuracy

The primary concern is kinetic realism, dynamic consistency, and interpretability—not merely held-out test performance.

Extrapolation Risk

A surrogate with good endpoint metrics may produce biologically implausible or temporally inaccurate predictions outside the training domain.

Regulatory Gap

Inadequate validation may lead to incorrect predictions precisely in the high-risk scenarios QSP models are designed to address.

Dynamic Systems

QSP models describe transient dynamics, feedback loops, and multi-phase responses. Surrogates must reproduce temporal features—not just steady-state endpoints. Timing errors in cytokine peaks (CAR-T) or reticulocyte nadirs (EPO) can be clinically significant.

Multiscale Stiffness

Biological systems span fast processes (receptor binding, seconds–minutes) and slow ones (disease progression, weeks–months), producing numerically stiff ODEs. Surrogates may inherit numerical artefacts or miss stiff transients if training data lack sufficient temporal resolution.

Parameter Identifiability

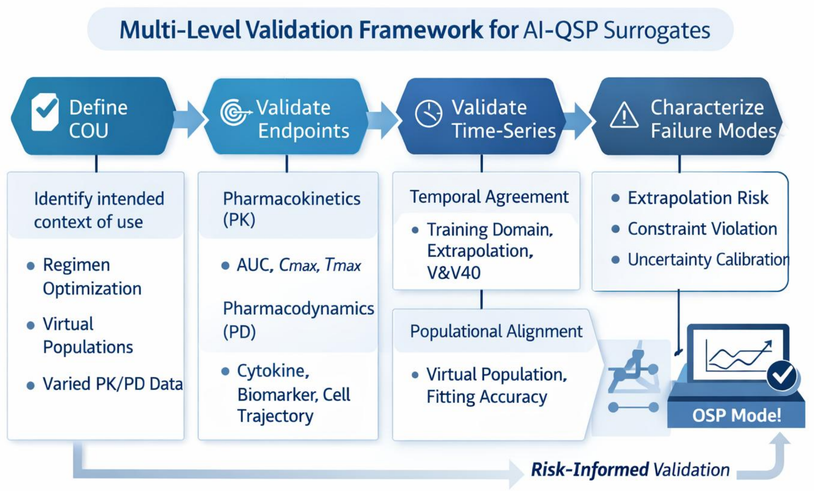
QSP models frequently exhibit *sloppy* parameter spaces where multiple configurations produce similar outputs. Surrogates may learn non-causal correlations, producing accurate interpolations but misleading extrapolations—a failure mode invisible to standard held-out metrics.

Biological Constraints

Mechanistic models respect positivity, physiological bounds, and conservation laws. Standard ML architectures impose no such constraints. Non-physical predictions—negative concentrations of cytokines etc—must be detected and treated as model failure indicators.

The appropriate surrogate class depends on the context of use, training budget, model dimensionality, and required interpretability. Table 1 summarises the principal classes.

Surrogate Class	Core Idea	Training Data	Strengths	Common Failure Modes	Regulatory Positioning
Black-box ML (MLP, RF, XGBoost)	Full input→output mapping	High (10^3–10^4+)	Fast inference; strong interpolation	Poor extrapolation; constraint violations	Exploratory COUs only
Gaussian Process / Kriging	Probabilistic kernel interpolation	Low–moderate (100s–1,000s)	Calibrated uncertainty; data efficient	Poor scaling to high-d	Often well-received
PINNs	Governing equations as loss penalties	Moderate–high	Physical consistency	Training instability with stiff systems	Needs careful justification
UDEs / Neural ODE hybrids	Neural net in unknown ODE components only	Low–moderate	Excellent inductive bias; good extrapolation	Solver sensitivity	Strong regulatory story
Residual hybrid surrogate	ML learns correction to reduced mechanistic model	Low	Interpretable error structure; data-light	Depends on base model quality	Strong if base model qualified
Latent-space dynamical surrogate	Compress trajectories; emulate latent dynamics	Moderate	Handles high-dimensional outputs	Latent misinterpretation; decoder bias	Acceptable if interpretability demonstrated



The Eight-Step Validation Workflow

The workflow begins with COU definition (Step 1) and simulation campaign design (Step 2), followed by progressive validation tiers: endpoint accuracy (Step 4), trajectory fidelity (Step 5), uncertainty calibration (Step 6), and distributional agreement (Step 7), with baseline comparison at each tier (Step 3). Step 8 addresses biological plausibility and computational efficiency. Failure-mode diagnostics are applied throughout.

Failure Modes & Diagnostics

Extrapolation Failure

Most common failure mode. Detect via boundary stress tests and Mahalanobis distance to training data. Restrict surrogate to defined domain of applicability or adopt hybrid architectures.

Constraint Violations

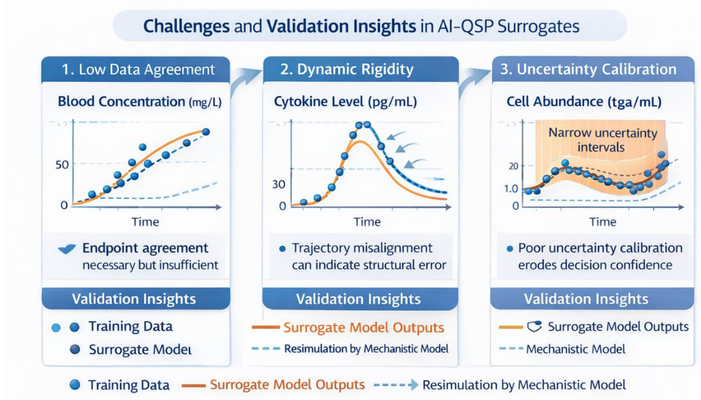
Negative concentrations, mass-balance violations. Mitigate via softplus transformations, constraint-penalty loss terms, or mechanistic scaffolding.

Temporal Mismatch

Correct AUC but biased Cmax/Tmax or phase-shifted trajectories. Detect via residual-vs-time plots and landmark error analysis.

Uncertainty Miscalibration

Overconfident intervals failing to capture mechanistic variability. Assess on held-out simulations; report coverage probabilities as part of validation evidence.



Acceptable Error Thresholds by Context of Use

Acceptable surrogate error is inherently context-dependent and must be stated explicitly. These are illustrative guidance values, not regulatory acceptance criteria.

Context of Use	Illustrative Acceptable Error	Rationale
Exploratory screening	10–20% endpoint error	Ranking more important than precision
Dose/regimen comparison	≤10% on AUC/Cmax	Preserves relative ordering of regimens
Virtual population generation	KS distance ≤ 0.10–0.20	Distributional fidelity critical
Decision-critical support	≤5% endpoints + trajectory fidelity	High risk of incorrect decisions
Regulatory-facing analyses	Conservative, COU-specific thresholds	Align with risk-informed credibility frameworks

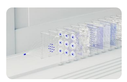
When diagnostics indicate surrogate failure, best practice is to suspend surrogate use for the affected COU, revert to the mechanistic model, expand training data in the failing region, and document all failure modes, diagnostics, and mitigation steps.

Hybrid Mechanistic–AI Surrogate Approaches

Hybrid approaches combine the extrapolative reliability of physics-based formulations with the representational flexibility of data-driven learning. They are particularly advantageous when training data are sparse, extrapolation is unavoidable, biological constraints are critical, or trajectory fidelity matters more than endpoint accuracy.

PINNs

Governing equations encoded as loss penalties. Most useful when equations are well-specified but expensive to solve. Training instability for stiff pharmacological systems limits current regulatory use.

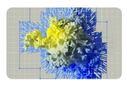


Universal Differential Equations

Neural networks replace only unknown ODE components; known structure is retained. Reduces learning dimensionality, exhibits superior extrapolation, and provides strong inductive bias.

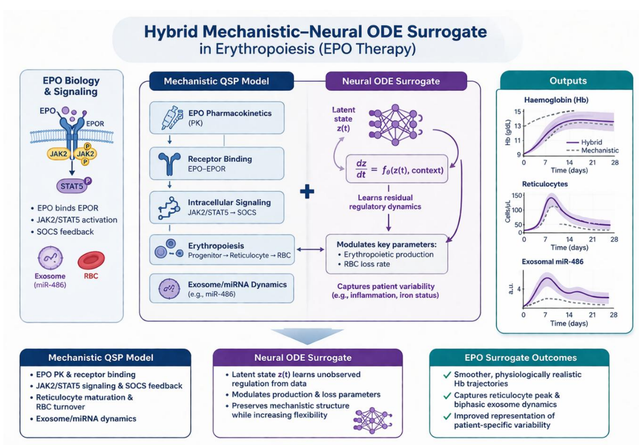
Residual Hybrid Models

ML learns only the correction to a reduced mechanistic model. Dramatically reduces surrogate complexity and provides clear interpretation of what the ML component adds beyond established biology.



Constrained Architectures

Biological constraints embedded directly into architecture or training objective (e.g., softplus for positivity, mass-balance parameterisation). Easier to train than PINNs; substantially reduces non-physical predictions.



Hybrid mechanistic–Neural ODE surrogate model for erythropoiesis under EPO therapy. A mechanistic QSP model describing EPO pharmacokinetics, JAK2/STAT5 signalling, and erythropoiesis is augmented with a Neural ODE latent state z(t) that learns residual regulatory dynamics from data. The latent state modulates biologically meaningful parameters, enabling improved representation of patient-specific variability while preserving mechanistic structure and biological constraints. Comparison panels show improved haemoglobin trajectory fidelity and reticulocyte peak reproduction relative to the mechanistic-only model.

Overarching Principles

COU is the anchor

Without a clearly articulated COU, it is impossible to determine what level of surrogate error is acceptable or which validation tiers are required.

Endpoint accuracy is insufficient

Trajectory fidelity must be assessed independently—surrogates frequently achieve acceptable endpoint agreement while misrepresenting clinically meaningful temporal dynamics.

Failure modes must be actively tested

Extrapolation failure, constraint violation, temporal mismatch, and uncertainty miscalibration are systematic, not random, and require targeted diagnostics.

Limitations & Future Directions

The peer-reviewed literature reporting multi-level surrogate validation in QSP settings remains sparse. Many published studies report endpoint accuracy without trajectory or uncertainty validation, and benchmark datasets for systematic comparison are largely absent.

Three Priority Directions

1. Publicly available benchmark datasets covering diverse QSP application domains
2. Further advancement of hybrid mechanistic–AI modelling approaches
3. Standardised validation reporting guidelines to enable cross-study comparison and facilitate regulatory acceptance

Conclusions

> "Surrogates do not replace mechanistic models; they scale their use—and the credibility of that scaling must be explicitly demonstrated."

Transformative Opportunity

AI-QSP surrogates enable large-scale GSA, VP generation, Bayesian estimation, and regimen optimisation that would otherwise be computationally prohibitive.

Rigorous & Actionable Framework

The eight-step workflow integrates COU definition, simulation design, baseline comparison, endpoint/trajectory/uncertainty/distributional validation, and failure-mode diagnostics into a coherent, risk-informed hierarchy.

Regulatory Grounding

Anchored in FDA CM&S guidance and ASME V&V40, the framework provides practical guidance for deploying AI surrogates as validated computational accelerators in both research and regulatory contexts.

*Correspondence: goryanin@gmail.com**Submitted to CPT. Available on BioRxiv and www.iqanov.org*